ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 17 за 2024 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

М.Ю. Кислицына, Ю.Н. Орлов

Статистический анализ полного корпуса художественной литературы на русском языке и распознавание автора

Рекомендуемая форма библиографической ссылки: Кислицына М.Ю., Орлов Ю.Н. Статистический анализ полного корпуса художественной литературы на русском языке и распознавание автора // Препринты ИПМ им. М.В.Келдыша. 2024. № 17. 24 с.
https://doi.org/10.20948/prepr-2024-17
https://library.keldysh.ru/preprint.asp?id=2024-17

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

М.Ю. Кислицына, Ю.Н. Орлов

Статистический анализ полного корпуса художественной литературы на русском языке и распознавание автора

Москва – 2024

Кислицына М.Ю., Орлов Ю.Н.

Статистический анализ полного корпуса художественной литературы на русском языке и распознавание автора

Собрана статистика эталонных триграмм для полного корпуса литературных текстов на русском языке, включая переводных зарубежных авторов. Построены распределения расстояний от отдельных текстов до эталонов. Протестирован метод ближайшего эталона для распознавания автора текста. Определена ошибка по жанрам, подгруппам авторов и по корпусу в целом. Проведена классификация ошибок для разработки метода коррекции.

Ключевые слова: триграммы, метод ближайшего соседа, распознавание автора текста

Kislitsyna M.Yu., Orlov Yu.N.

Statistical analysis of the complete corpus of fiction in Russian and recognition of the author

Statistics of reference trigrams for the complete corpus of literary texts in Russian, including translated foreign authors, have been collected. Distributions of distances from individual texts to standards are constructed. The nearest reference method for recognizing the author of the text has been tested. The error was determined by genres, subgroups of authors and by the corpus as a whole. A classification of errors has been carried out to develop a correction method.

Keywords: trigrams, nearest neighbor method, text author recognition

Содержание

Введение

В настоящей работе впервые анализируется максимально большой корпус литературных текстов на русском языке. Целью анализа является численное исследование статистических свойств полного, насколько это возможно на текущий момент, собрания сочинений всех писателей, в том числе и переведенных на русский язык. Под «всеми» писателями подразумеваются те авторы текстов художественной прозы, которые имеют не менее пяти произведений, каждое из которых содержит не менее 10 тысяч знаков без учета пробелов. Такое ограничение снизу обусловлено тем, что мы развиваем статистический подход к идентификации автора текста. В этой связи для сбора статистики требуется, чтобы у автора было несколько произведений достаточного объема, поскольку надо построить распределение частот символов с точностью, лучшей, чем различие между разными распределениями. Далее мы приведем необходимые количественные оценки.

Интерес к данной теме обусловлен распространением технологии искусственного интеллекта на различные области человеческой жизнедеятельности. Машинная обработка текстов на естественных языках представляет одно из важных направлений применения этой технологии для задач классификации текстов и массивов документов [1-3]. Существует большое количество компьютерных программ для их обработки. Все они используют методы статистики для автоматического определения языка, жанра или тематики, а также автора некоторого достаточно большого текста.

При этом применяются два основных подхода (см. [4, 5]). Первый состоит в экспертном отборе определенных текстов для обучения манере написания фраз, т.е. основан на мнении эксперта относительно правил написания произведения определенного формата. Ограничение этого подхода по точности связано с тем, что, во-первых, существуют разные экспертные мнения, и, во-вторых, автор может менять манеру письма.

Второй подход чисто статистический. В нем есть определенная претензия на объективность в том смысле, что методом может воспользоваться любой исследователь и, независимо от личной точки зрения, всегда будет получать один и тот же результат применительно к одному и тому же набору документов. В основном статистические методы основаны на анализе наиболее часто употребляемых слов, знаков препинания, букв, абзацев и иных чисто технических характеристик текста. Неполная объективность метода для задачи идентификации автора состоит в том, что, вообще говоря, автору нельзя запретить использовать те слова лексикона, которые использует и другой автор, а также те, значения которых автор не знал, но посмотрел в словаре и употребил в определенном контексте. В этом смысле статистика букв более объективна, чем статистика слов, но возможно, что она не обладает авторским своеобразием. Этот последний аспект и следует проверить.

Поскольку в настоящее время нет общепринятой «теории творчества», то все методы, использующие математический аппарат для получения каких-то результатов применительно к неизмеримым непосредственно понятиям, таким как жанр, авторство, мастерство и т.п., являются эвристическими. Следовательно, необходимо каждый такой метод тестировать на точность и устойчивость по отношению к параметрам выборки текстовых документов. С этим есть определенные трудности, поскольку для тестирования надо собрать максимально полный объем данных и провести их предобработку или разметку, чтобы проведение статистического эксперимента было корректным.

Здесь следует подчеркнуть, что имеющиеся многочисленные статьи на тему распознавания автора по тем или иным характеристикам текста хотя и кажутся статистически достоверными, все же оперируют с весьма малыми выборками авторов. Как правило, рассматривается число авторов порядка 10 (см. работы [6-8]), что не представляет «промышленного интереса». Дело в том, что для идентификации автора требуется статистически определить некоторый авторский эталон из тех элементов текста, на которых основан метод. Тогда отдельные произведения автора трактуются как выборка из эталона, что позволяет далее применить метод ближайшего соседа для определения того эталона, который ближе всех к данному тексту. Однако в работе [9] было показано, что задача идентификации выборки среди системы эталонных функций распределения зависит от конкретной системы эталонов, а не только от расстояний между ними. Поэтому, например, две группы по 10 эталонов с одним и тем же распределением расстояний между ними могут обладать различными идентифицирующими свойствами. Следовательно, каждый набор текстов обладает, строго говоря, только ему свойственной ошибкой распознавания автора, которая не может быть распространена на другие аналогичные по объемам группы текстов и авторов. В этом плане эксперименты с малым количеством «образцов» проводятся исключительно с иллюстративной целью, а именно: показать, что существует непустое множество объектов, для которых предлагаемый метод дает приемлемые в статистическом смысле результаты. Однако собственно научной ценности – ни в области математической статистики, ни в конкретной предметной области – такие результаты не имеют. Для получения таковых необходимо рассмотреть настолько большой корпус эталонов, который в рамках доступной точности проведения эксперимента не чувствителен к удалению любого из эталонов.

В некоторых работах представлены результаты более развернутых экспериментов, хотя и они проводятся для весьма неполных корпусов литературных текстов. Так, в статье [10] анализируется набор текстов, написанных 50 авторами – классиками русской литературы, но текстов относительно немного, всего 215. В статье [11] рассмотрено распознавание авторства более чем 1700 текстов среди 100 писателей, что хотя и более представительно, чем в других работах, но все же мало по сравнению с полным списком литературы на русском языке.

Кроме того, в исследованиях с небольшим числом авторов имеется скрытая тенденциозность их подбора. Естественно, что сначала есть желание

собрать тексты известных писателей, таких произведений относительно немного. Затем для разнообразия к ним можно добавить романы нескольких менее популярных авторов, после чего активность исследователя по составлению корпуса спадает. В результате получается, что писатели с весьма индивидуальным стилем прекрасно отделяются один от другого, ошибка оказывается невелика. Кроме того, в упомянутой выше работе [11] отмечен важный для понимания самого процесса тестирования результат: если выбросить ошибочно определенные тексты и соответствующих плохо определяемых авторов, то оставшиеся примерно 1,5 тысячи текстов 90 авторов опознаются правильно. Следовательно, всегда есть возможность так собрать тексты и авторов, что они идентифицируются безошибочно, но это не может служить основанием для утверждения о том, как устроена модель литературного творчества, а показывает лишь, что применяемый метод адекватен в том смысле, что имеет достаточно большое множество примеров правильного своего применения. Поэтому полнота корпуса играет принципиальную роль для понимания зависимости точности метода от размеров и состава корпуса. Одно дело – распознать, допустим, одно из десяти крупных произведений Тургенева среди ста текстов десяти других русских классиков, а другое – один из небольших рассказов Тургенева среди 10 тысяч других авторов с числом произведений под 100 тысяч. Таких вычислительных экспериментов до настоящего времени проведено не было. Хотя электронные библиотеки сейчас вполне доступны, проблема масштабного анализа текстов связана с существенным преобразованием структуры библиотеки, о чем будет сказано далее.

В обзорной работе [12] сравнивается эффективность чисто статистических методов анализа, основанных на подсчете формальных показателей, таких как число букв, слов, знаков препинания и т.п., с экспертными методами анализа авторского стиля, оборотов речи, использования литературных приемов. Авторы [12] приходят к выводу, что, хотя для литературоведов более ценен экспертный метод, он не обладает достаточной точностью на большом корпусе текстов и, что более существенно, не может быть адекватно реализован в виде формальной компьютерной программы. В то же время статистика букв или буквосочетаний хотя и не имеет непосредственного литературного смысла, может быть вполне однозначно сопоставлена каждому тексту с указанием погрешности в рамках формальных критериев. Тем самым в контексте задачи машинного распознавания атрибутов текстов статистический метод более эффективен, т.е. имеет меньшую ошибку, чем экспертный.

Задача автоматической классификации текста по атрибутам является весьма актуальной в области развития информационных технологий и анализа больших данных. Кроме того, она находится на стыке наук: математической статистики и лингвистики, что может способствовать развитию обеих отраслей знания. Однако многочисленные методы машинного обучения (см., напр., [13-16]), используемые для той или иной обработки и анализа текстовых документов, представляют собой «черный ящик» не только для лингвистов, но и собственно для математиков. Не ясно, насколько эти методы, настроенные на

решение задачи классификации в рамках конкретно отобранного корпуса текстов, робастны применительно к другим выборкам и другим задачам. Кроме того, в силу специфики настройки сложно предложить процедуру ее коррекции.

В нашей работе алгоритм обучения, т.е. сбора статистики, абсолютно прозрачен. Поскольку мы не занимаемся задачей кластеризации, а кластеры формируются заранее, то основная задача – собрать возможно более точный эталон кластера. В качестве эталона мы используем распределение эмпирических частот триграмм, т.е. троек подряд идущих буквосочетаний без пробелов и без различения строчных и прописных букв. Цифры и знаки препинания игнорируются. Каждому автору, который считается известным точно, сопоставляется вектор триграмм на основе анализа всех его произведений. Затем методом кросс-валидации проводится опознавание автора текста среди известных эталонов. Для сравнения мы приводим также результаты аналогичной процедуры для однобуквенных и двухбуквенных авторских эталонов.

1. Методика проведения эксперимента

Для анализа была использована электронная библиотека текстов на русском языке, содержащая 630 563 произведения 122 832 авторов. Поскольку в библиотеке содержатся именно книги, а мы планируем анализировать только отдельные произведения, каждое из которых написано одним автором (в соответствии с тем, как указано на титульном листе), то возникает большая техническая проблема преобразовать такую библиотеку именно в корпус текстов с классификацией по интересующим нас атрибутам. Решение этой проблемы потребовало значительных усилий как на уровне составления просеивающей программы, так и на уровне модели просеивания, поэтому мы сочли необходимым описать этот первый этап подготовки корпуса достаточно подробно.

Этапы преобразования библиотеки.

1. Сначала исключаются все тексты, написанные в соавторстве или коллективом авторов.

2. Затем отбираются авторы, имеющие не менее пяти текстов, каждый из которых имеет объем не менее 10 тыс. знаков без учета пробелов.

3. На следующем этапе из рассмотрения исключаются словари, справочники, научно-техническая и иная специальная литература (садоводство, кулинария, медицина, математика, здоровый образ жизни и т.п.) – все, что не является литературным текстом. Кроме того, из корпуса удалялись все тексты религиозного значения (канонические тексты, молитвы, практики и т.п.), поскольку в таких произведениях обычно содержится множество цитирований.

4. Поскольку мы не анализируем тексты, в которых буквосочетания возникают в силу специальных ограничений, внешних по отношению к смыслу произведения, то исключаются стихи, тексты в виде палиндромов и т.п. формы.

5. Из рассмотрения исключаются пьесы, так как в них также содержится навязанная форма – многократное упоминание действующего лица.

6. Некоторые произведения содержатся в библиотеке в нескольких экземплярах, не всегда идентичных друг другу. Бывает так, что в отдельных сборниках тексты представлены не полностью, а в сокращенном варианте. В таких случаях фрагменты идентифицируются и оставляется самый длинный вариант.

7. Один и тот же автор может писать под разными псевдонимами, что требует проверки существования такого писателя. Тексты «разных», но по факту совпадающих авторов после идентификации приписываются только одному имени.

8. Переводные тексты могут совпадать, но иметь различные заголовки. Такое встречается, например, тогда, когда в книге представлены два романа или повести одного зарубежного писателя, а книга позиционируется в библиотеке по названию на обложке, которое может вообще не соотноситься с текстами внутри. Кроме того, одно и то же произведение может быть несколько раз переведено разными переводчиками. Среди подобных переводных текстов в корпусе оставлялось произведение с наибольшей длиной.

9. Буквы е и ё отождествляем, поскольку в разных произведениях умляут в одних и тех же словах может использоваться, а может и опускаться.

В результате проведенной подготовки корпуса текстов библиотека сократилась до 108 518 текстов, которые написаны 8 287 писателями. Из них 5 084 автора (61% авторов) – русскоязычные, остальные 3 203 автора (39%) рассматриваются в переводах. По количеству литературных произведений состав корпуса следующий: имеется 66 215 текстов на изначально русском языке, что также составляет 61% от общего числа текстов корпуса, а 42 303 текста – переводные. Полный список авторов и текстов корпуса представлен в [17]. Далее именно этот набор произведений будет называться «Полный Корпус» или просто корпус.

Общий объем анализируемых произведений составляет 34 395 575 697 (примерно 34 млрд) знаков, из них 20 млрд знаков соответствуют текстам на изначально русском языке. Таким образом, русскоязычные и переводные произведения как по суммарной длине, так и по количеству текстов и количеству авторов соотносятся примерно как 60:40. Для всей мировой литературы такое соотношение может показаться странным, но надо учесть, что в последнее время активно стали появляться «электронные коммерческие писатели» в виде самиздата, тогда как не каждый зарубежный писатель переводится профессиональным переводчиком на русский язык.

Оценка точности эмпирических частот символов.

Оценим точность, с которой определяется вероятность нахождения некоторого символа в тексте из n знаков. Предположительно, каждый автор характеризуется своим стационарным распределением таких вероятностей. Обозначим через g_j^0 теоретическую вероятность, характерную для данного автора, появления в тексте символа j. Под символом понимаем некоторое

буквосочетание. Пусть при анализе текста определена эмпирическая частота этого символа, равная g_j. Эмпирическая дисперсия наблюдаемых частот равна $\sigma_j^2 = g_j\left(1-g_j\right)$. Поскольку отклонения от среднего значения стационарной генеральной совокупности распределены асимптотически нормально, то интервальная оценка среднего значения g_j^0 нормального распределения на уровне значимости α по выборке длины N определяется формулой (см., напр., [18], стр. 319):

$$\left|g_j^0 - g_j\right| \le q_{1-\alpha/2}^t(N-1)\frac{\sigma_j}{\sqrt{N}}, \tag{1.1}$$

где $q_{1-\alpha/2}^t(N-1)$ есть квантиль распределения Стьюдента с $N-1$ степенями свободы. Так как $N >> 1$, то вместо квантиля распределения Стьюдента можно использовать квантиль $u_{1-\alpha/2}$ стандартного нормального распределения. Тогда на уровне значимости α ширина доверительного интервала оценивается величиной

$$\left|g_j^0 - g_j\right| \le u_{1-\alpha/2}\sqrt{\frac{g_j\left(1-g_j\right)}{N}}. \tag{1.2}$$

Потребуем теперь, чтобы отношение длины доверительного интервала к оцениваемому значению было равно уровню значимости интервала. Для такой согласованной оценки α получаем из (1.2):

$$\frac{\alpha}{u_{1-\alpha/2}} \le \sqrt{\frac{1-g_1}{Ng_1}}. \tag{1.3}$$

Для численных оценок удобно воспользоваться аппроксимацией Смирнова (см. [19], стр. 27):

$$u_{1-\alpha/2} \approx \sqrt{-\frac{\pi}{2}\ln\left(1-\left(1-\alpha\right)^2\right)}. \tag{1.4}$$

Например, для уровня значимости 0,05 левая часть формулы (1.3) равна приблизительно 0,025. Поэтому достаточная выборка, по которой можно оценить генеральную вероятность g_j^0 с указанной точностью, составляет примерно $N \approx 1600/g_j^0$, что для максимальной частоты триграмм равно примерно 1 млн, а для минимальных ненулевых частот – порядка 10 млрд. Последнее, разумеется, невыполнимое требование для эталона отдельного автора. Однако каждую частоту нет необходимости оценивать с одной и той же точностью, поскольку малые частоты вносят в нормировку вклад, много меньший указанной точности, и потому ошибка в их оценке может быть не 5 %, а хоть 500 %. В этой связи удобно использовать вместо условия (1.3) некоторую средневзвешенную точность, введенную в [5]:

$$\frac{u_{1-\varepsilon/2}}{\varepsilon} \le \frac{\sqrt{N}}{\Sigma_N(n)}, \tag{1.5}$$

где

$$\Sigma_N(n) = \sum_{j=1}^{K} s(j;N) = \sum_{i=1}^{K} \sqrt{f^{(n)}(j;N) \cdot \left(1 - f^{(n)}(j;N)\right)}, \qquad (1.6)$$

K есть количество символов, а n – длина буквосочетания. При заданной точности ε формула (1.5) для знака равенства дает оценку на минимальную длину выборки, при которой эта точность достигается на распределении частот в норме L1. В этом смысле значимость оценки эталонного распределения 0,05 достигается на текстах суммарной длины примерно 4 млн знаков, что по большей части выполнимо для одного автора. Оно, однако, не выполняется практически ни для одного отдельного текста, что формально препятствует распознаванию авторов текстов данным методом. Но надо учесть, что оценки сделаны в предположении независимости появления символов в тексте, что заведомо не так. Поэтому приведенные выше формулы не дают ответа на вопрос, на какой длине текста распределение символов становится стационарным. Ответ может быть получен лишь численно. В [5] эта длина названа длиной представительности автора. На ряде примеров было показано, что эта длина порядка 10 тыс. знаков. Поэтому, в частности, в данном эксперименте мы и ввели такое ограничение снизу на длину текста.

Заметим также, что с увеличением порядка n-граммы достоверность оценивания компонент вектора частот снижается, поскольку число символов N в книгах авторов ограничено и в среднем равно 400 тыс. Естественное условие $K < N$, необходимое для статистически корректного оценивания частот при распределении N значений по K ячейкам, приводит на практике к ограничению использования n-грамм числом $n = 3$.

Метод идентификации автора.

Итак, пусть определено эмпирическое распределение $D_a^i(j)$ частот символа j в i-м тексте автора a, и N_a^i есть число символов в данном i-м тексте. Пусть n_a есть число произведений автора a. Тогда эмпирической оценкой эталона $F_a(j)$ автора a является взвешенное распределение частот по совокупности всех текстов, достоверно принадлежащих данному автору:

$$F_a(j) = \frac{1}{N_a} \sum_{i=1}^{n_a} N_a^i D_a^i(j), \quad N_a = \sum_{i=1}^{n_a} N_a^i . \qquad (1.7)$$

Расстояние между текстами, а также между текстом и эталоном будем понимать в смысле расстояний между соответствующими распределениями в норме L1. В силу разного количества символов в разных текстах такое сравнение не вполне корректно в статистическом смысле. Тем не менее, как было показано в [5], тексты одного автора длиной более 10 тыс. символов имеют распределение, отличающееся от эталона на величину, примерно в два раза меньшую, чем между собой. Поэтому такое сравнение вполне адекватно.

Метод кросс-валидации состоит в том, что на момент сравнения данного текста с эталонами корпуса этот текст исключается из составляющих эталона данного автора. С учетом этого замечания расстояние от текста до эталона корпуса дается формулой [11]:

$$z_{ab}^{i} = \frac{1}{1-\delta_{ab} N_b^i / N_b} \sum_{j=1}^{J} \left| D_a^i(j) - F_b(j) \right|, \qquad (1.8)$$

где δ_{ab} есть символ Кронекера. Формально для безошибочного распознавания автора текста требуется, чтобы

$$\forall i,a,b:\ z_{aa}^{i} < z_{ab}^{i},\ b \neq a\,. \qquad (1.9)$$

Условие (1.9) означает, что совместное распределение расстояний от текста до своего и до чужого эталонов должно иметь треугольный носитель. Тестирование корпуса текстов на предмет выполнения условия (1.9) составляет основную задачу данной работы.

Ошибкой идентификации будем называть долю текстов, автор которых был неверно определен по формуле (1.8). Формально это есть отношение числа нарушений неравенства (1.9) к общему числу текстов.

2. Статистическое описание корпуса текстов

Ниже на рис. 1-8 представлены основные статистические характеристики анализируемого корпуса литературных произведений.

Многочисленные тематические разделы электронной библиотеки были агрегированы в следующие 10 основных жанровых классов (рис. 1).

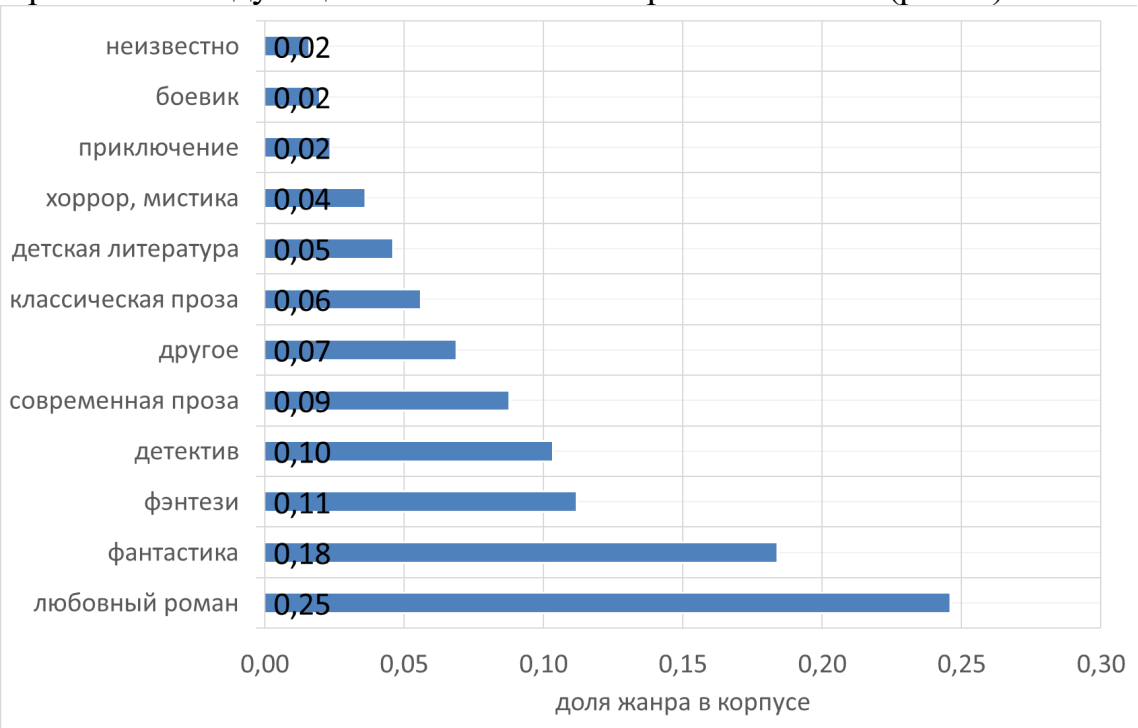

Рис. 1. Распределение текстов по жанрам

Из этого распределения следует, что собственно литературы как таковой (современная и классическая проза) имеется только 15% от общего числа произведений. Наибольшее число произведений (более 25 тыс.) относится к жанру «любовный роман», а наименьшее (порядка 2 тыс.) – к жанру «боевик». Не обсуждая социально-психологические аспекты такого положения дел в издании текстов, рискнем предположить, что отчасти это обусловлено типом

целевой группы, поскольку произведения данных жанров вызваны к жизни не авторским предложением, а читательским спросом. По-видимому, люди, которым важна тема стрельбы, книг особенно не читают, в отличие от сентиментальных дам.

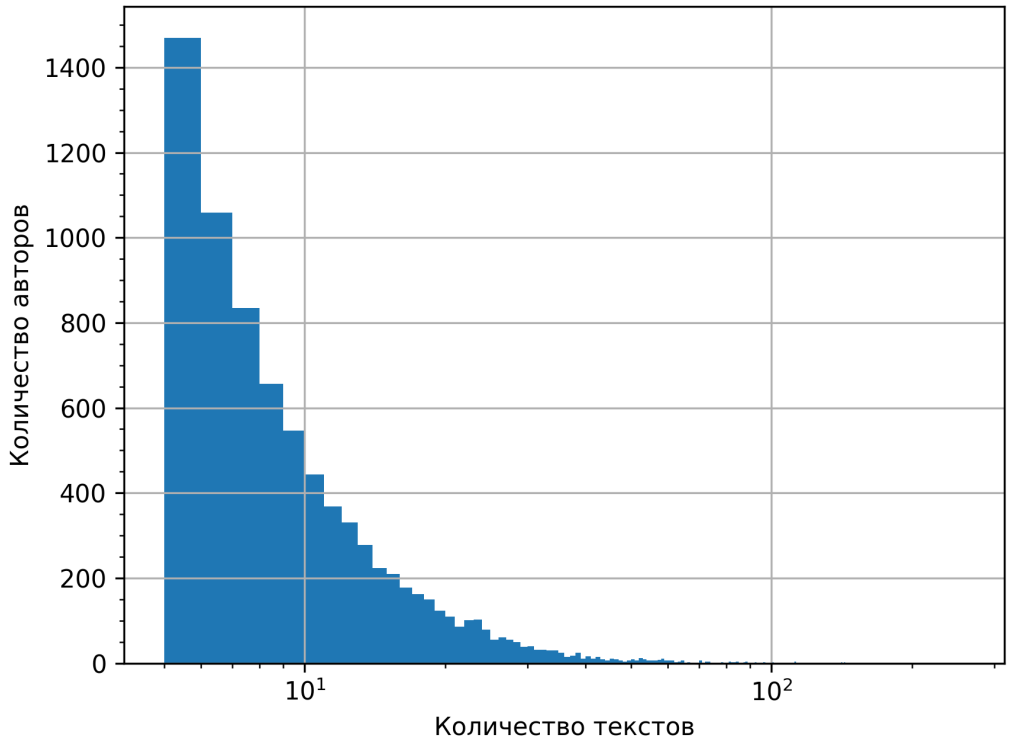

Рис. 2. Распределение авторов по числу текстов

На рис. 2 показано количество текстов корпуса в логарифмическом масштабе и соответствующее количество авторов. Первое значение – минимальное, равное пяти. Максимальное число текстов равно 374, автор – Барбара Картленд с огромным числом любовных романов. Заметим, что из них только 27 было идентифицировано ошибочно, что составляет всего 7% несмотря на то, что это переводная литература. Выпишем первую десятку «графоманов». Следом идут: Сергей Зверев (327 текстов, 160 ошибок (!)), Дарья Донцова (266, 2 ошибки), Роберт Стайн (256, 44 ошибки), Фридрих Незнанский (241, 61 ошибка), Владимир Колычев (225, 4 ошибки), Дарья Калинина (217, 1 ошибка), Александр Тамоников (214, 99 ошибок), Елена Арсеньева (205, 83 ошибки), Чингиз Абдуллаев (183, 9 ошибок). Структура ошибок анализируется далее в разделе 4.

Видно, что распределение авторов по числу текстов в интервале от 5 до 50 произведений в целом монотонно за исключением слабого скачка на уровне 22 текстов. С детерминацией 0,99 число авторов зависит обратно пропорционально от квадрата числа произведений.

На рис. 3 показано распределение текстов корпуса по числу знаков. Минимальное число знаков – 12 тыс., максимальное – примерно 1,5 млн.

Примерно 75 % текстов имеют длину менее 400 тыс. знаков, а все произведения больших объемов составляют содержимое последнего квартиля. Если в первых трех квартилях распределение по длине текста «примерно равномерно», то в последнем квартиле оно убывает как обратная длина текста в степени 7/2 (детерминация этой эмпирической зависимости также 0,99).

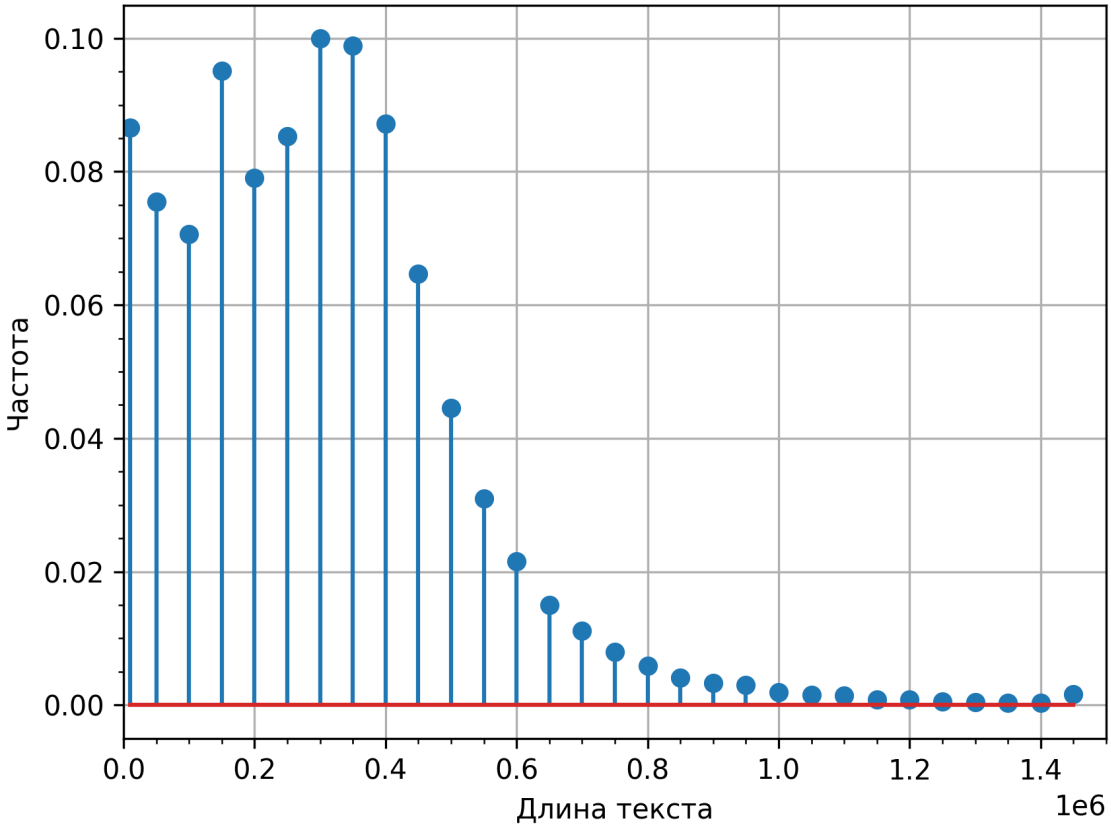

Рис. 3. Распределение текстов по числу знаков

Таким образом, среднее число произведений на одного автора составило 13, медиана этого распределения равна 9. Средняя длина произведения равна 317000 знаков. Число произведений с потенциально плохой идентификацией методом триграмм (их длина менее 30 тыс. знаков) составило 5028 или 4,6 % от общего числа текстов.

На рис. 4 показано распределение средней длины текста автора.

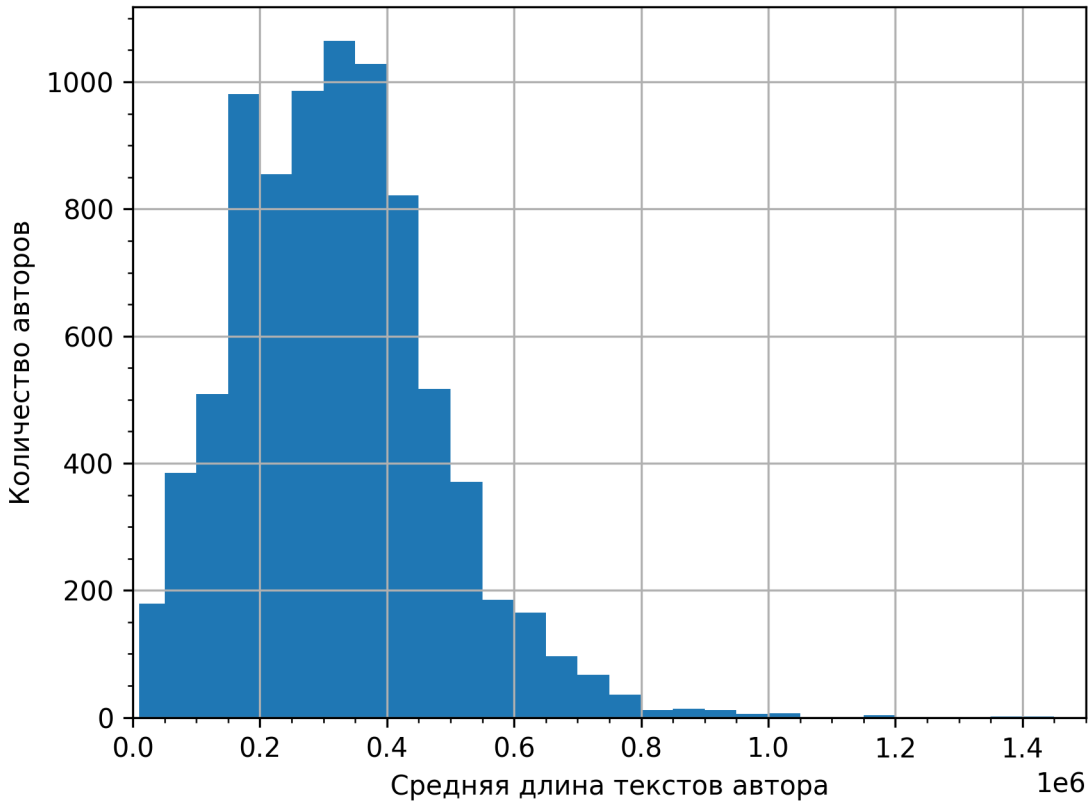

Рис. 4. Распределение средней длины текста автора

На рис. 5 и рис. 6 показаны распределения расстояний в норме L1 для триграмм – эталонов и отдельных текстов. Характерное расстояние между эталонами равно 0,3, а между отдельными текстами – примерно 0,5. При этом минимальные расстояния для обеих групп примерно совпадают и равны 0,25. Следовательно, если сравнивать между собой только тексты, а не близость между текстом и эталоном, то ближайший текст может быть совсем не того автора, что ближайший эталон. Этим объясняется выбранный нами метод ближайшего соседа.

Каждый текст и, соответственно, эталон, представленный в виде вектора частот триграмм, является вектором в пространстве размерности 32^3. Компоненты этого вектора неотрицательны, и их сумма равна единице, т.е. вектор определяет некоторую точку на 32768-мерном симплексе. Фактически его размерность существенно меньше, поскольку ненулевых компонент у этого вектора порядка 10 тыс., а с учетом объединения носителя для всех авторов – 11 тыс.

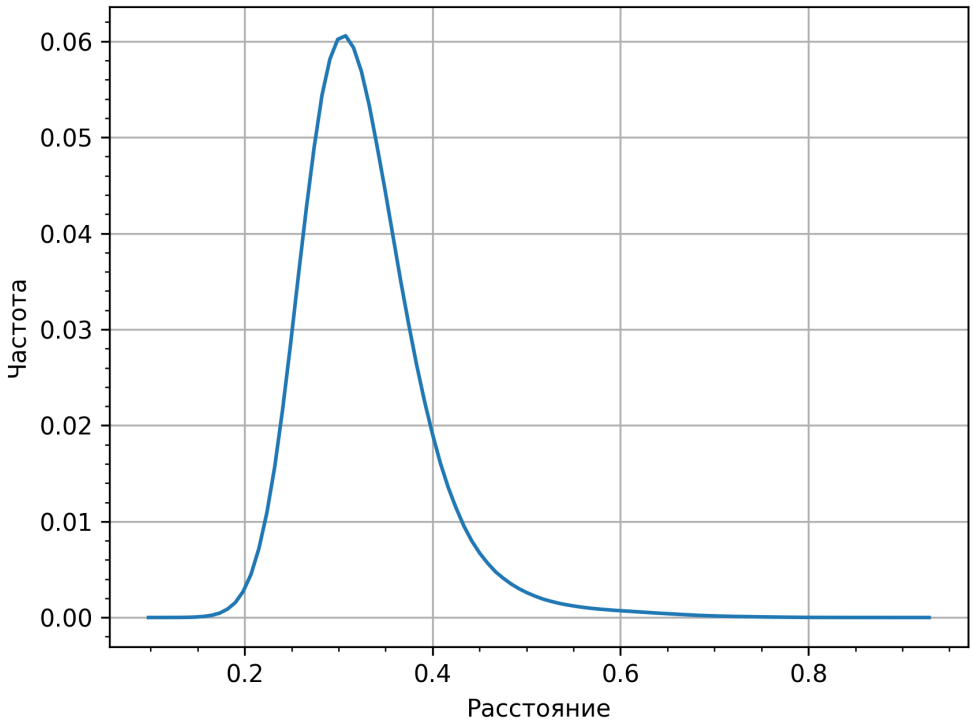

Рис. 5. Распределение расстояний в L1 между эталонами для n = 3

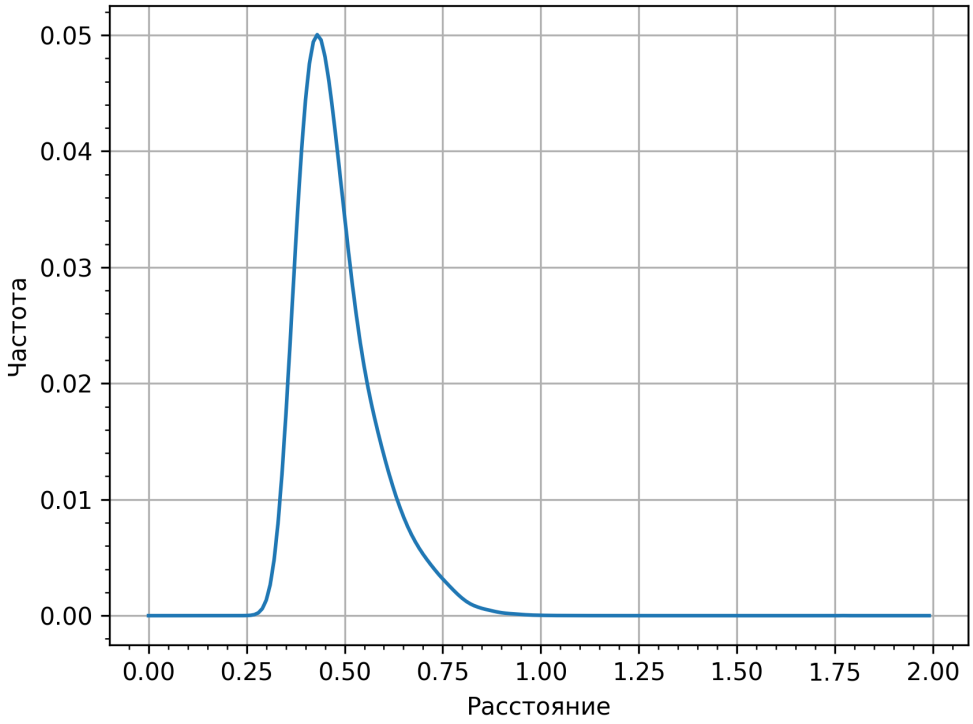

Рис. 6. Распределение расстояний в L1 между текстами для n = 3

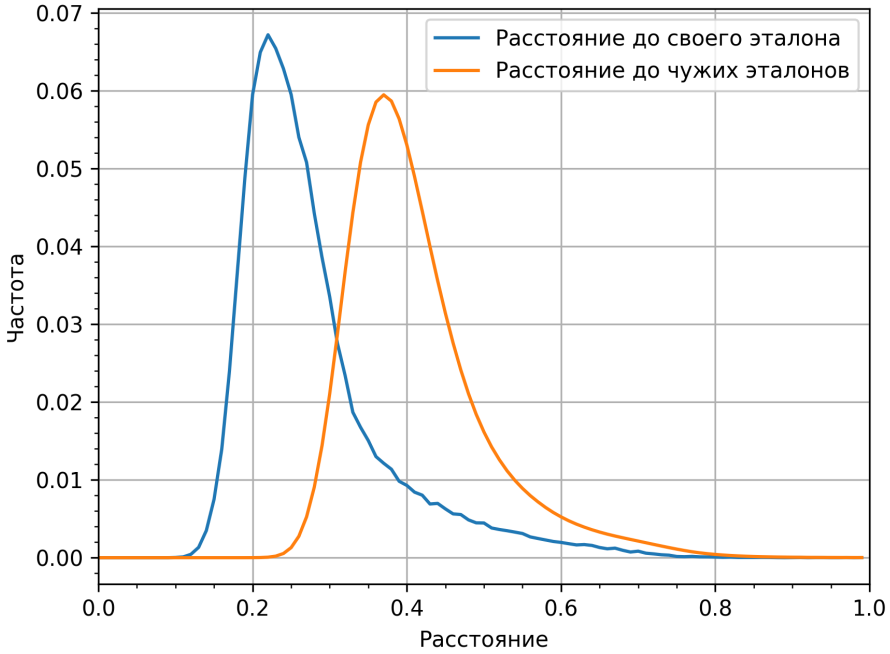

Рис. 7. Распределение расстояний в L1 между текстами и эталонами для n=3

Рассмотрим также величину нормированного размаха частоты для некоторого автора. Пусть эталонное значение некоторой частоты равно f. По текстам данного автора определяется минимальное значение частоты этого символа fmin и максимальное fmax. Для каждого автора вычисляется величина r=(fmax-fmin)/f.

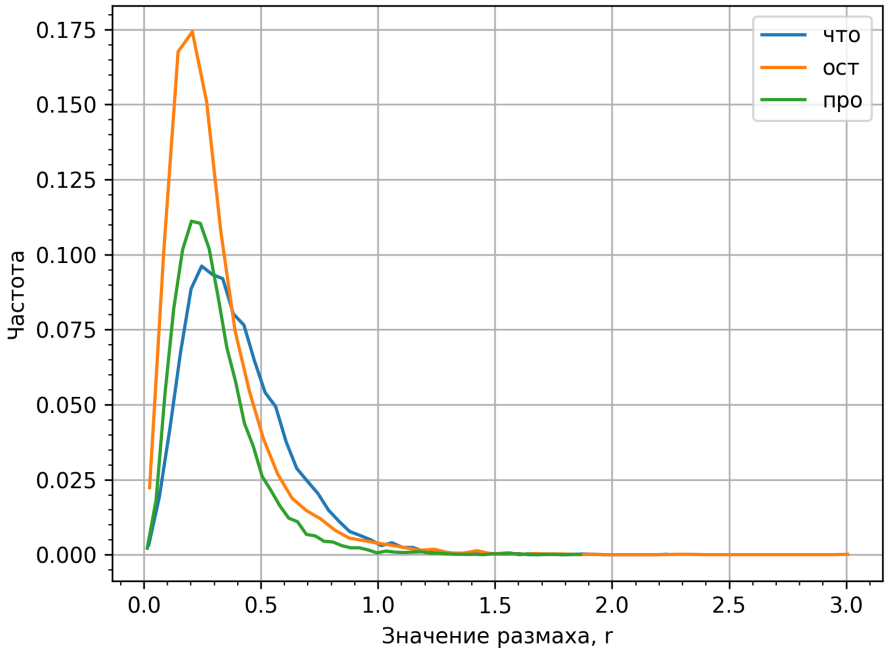

Рис. 8. Распределение величины нормированного размаха для трех наибольших частот эталона «что», «ост» и «про»

Из последних двух графиков можно сделать вывод относительно того, насколько плотно упакованы векторы текстов и эталонов на симплексе. Типичный размах составляет примерно 0,3 от частоты, 90% распределения набирается на промежутке от 0 до 0,5. Аналогичный размах этих же частот между разными авторами имеет характерную величину 0,6. Это показывает, что вариация распределений триграмм по текстам одного автора в целом в два раза меньше, чем между разными эталонами, чем и обусловлена возможность идентификации автора по распределению триграмм.

3. Результаты идентификации авторов

Основной результат статистического эксперимента по распознаванию автора текста содержится в таблице 1.

Таблица 1. Ошибка определения автора текста методом ближайшего эталона

n-грамма	n=1	n=2	n=3
ошибка, %	73,1	34,2	23,8

Выше мы отмечали, что потенциально «плохих» текстов, имеющих малую длину, в нашем корпусе 4,6%. Поскольку ошибка в 23,8% существенно больше, она не связана с этими текстами, а определяется другими причинами.

На рис. 9 приведено совместное распределение «свой-чужой». Это распределение заметно сдвинуто вправо, что объясняется тем, что большинство эталонов «чужих» авторов находится значительно дальше от эталона реального автора текста.

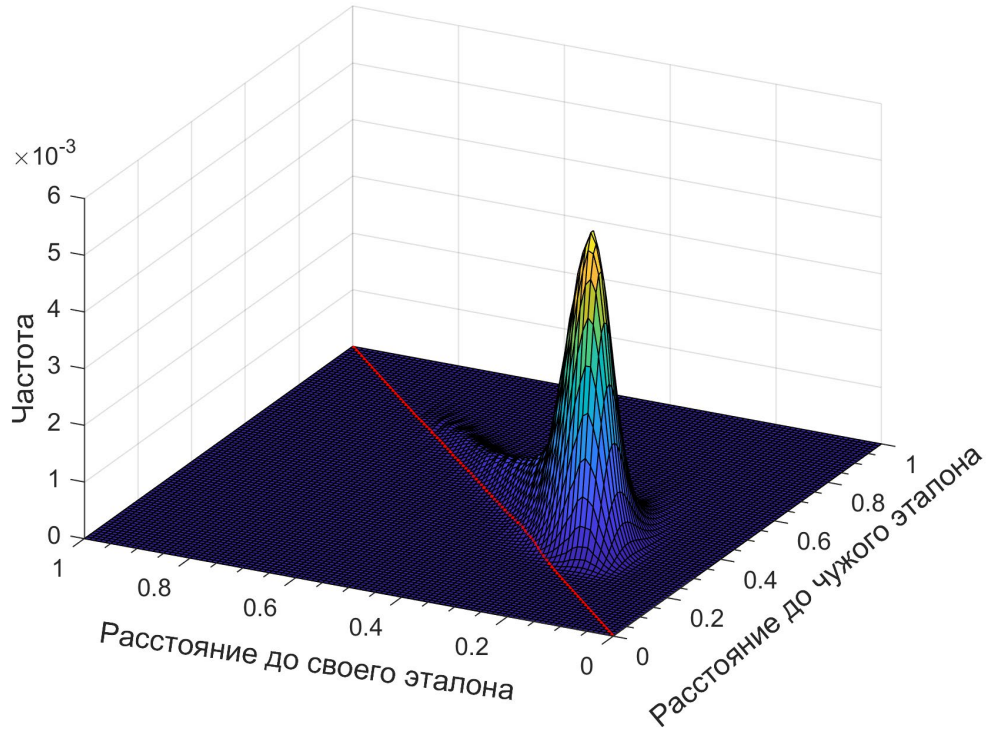

Рис. 9. Распределение расстояний «свой-чужой» для текстов корпуса для $n = 3$

На рис. 10 показана вероятность того, что ближайшим является расстояние до «своего» эталона.

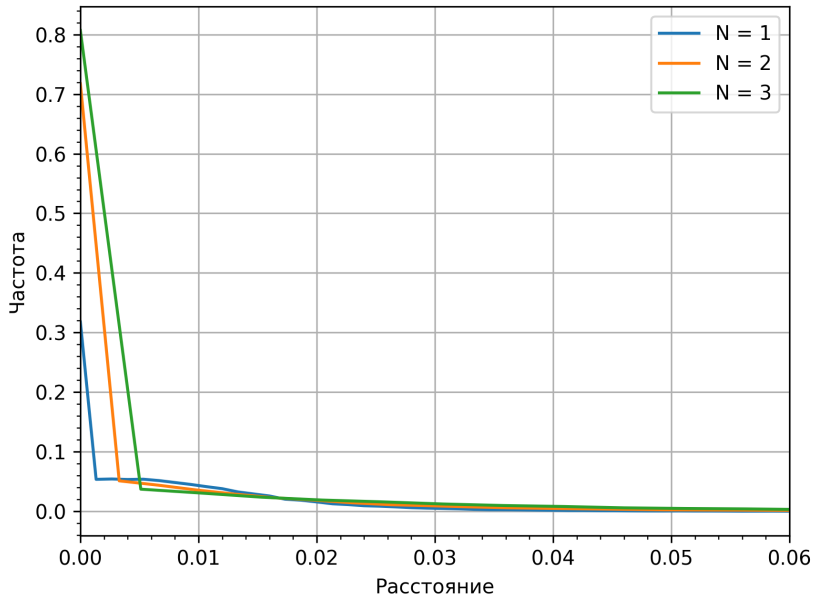

Рис. 10. Вероятность того, что ближайшим является «свой» эталон

Интересно проанализировать, зависит ли ошибка распознавания от жанра. Для этого мы рассмотрели идентификацию автора методом триграмм только внутри подкорпуса. Результаты приведены в таблице 2.

Таблица 2. Ошибка идентификации по жанрам

Жанр	Число авторов	Число текстов	Ошибка
фантастика	2353	26961	0,16
любовный роман	1567	18612	0,26
фэнтези	1132	11090	0,09
современная проза	966	8411	0,27
детектив	755	11507	0,10
детская	407	4349	0,17
классическая	372	3825	0,26
мистика	341	3414	0,20
боевик	200	2925	0,13
приключения	193	1881	0,10
философия	71	631	0,13
сетевая	185	1388	0,16

Таким образом, подтверждается результат работы [9] о том, что ошибка распознавания выборки среди системы эталонов зависит в основном от взаимного расположения эталонов. Если эталонов больше 10, а в нашем случае это число авторов, то ошибка не чувствительна ни к числу эталонов, ни к числу выборок. В силу этого мы не считаем, что ошибка в таблице 2 типична для

жанра, т.е. «ошибка в области классической литературы в два раза больше, чем в боевиках». Всегда можно подобрать подкорпус классиков с малой или вовсе с нулевой ошибкой, что неоднократно было показано в различных статьях.

Таблица 3. Ошибка для алфавитно упорядоченных групп авторов (по фамилии)

буква	число авторов	число текстов	ошибка
К	958	13296	0,15
С	715	9658	0,14
Б	683	8693	0,16
М	663	8432	0,15
Г	517	6857	0,14
Л	459	6134	0,14
Д	422	5670	0,14
П	403	5280	0,13
В	400	5055	0,14
А	397	5517	0,13
Р	370	4851	0,13
Т	265	3243	0,12
Ш	246	3034	0,11
Х	237	3137	0,14
Ф	222	2419	0,13
Н	200	2836	0,10
У	159	2236	0,15
О	156	1884	0,12
З	139	1952	0,08
Ч	136	1962	0,10
И	127	1593	0,10
Э	101	1087	0,09
Е	94	1095	0,08
Я	65	635	0,09
Ж	56	633	0,10
Ю	40	506	0,04
Щ	23	456	0,05
Ц	19	224	0,08
Й	8	71	0,04

Из таблицы 3 следует, что с увеличением числа авторов ошибка, вообще говоря, возрастает. Если эталонов меньше 100, можно ожидать ошибку порядка 0,05 – 0,10. Если от 100 до 1000, то ошибка порядка 0,15. Если авторов больше, как в полном корпусе, то ошибка 0,20 – 0,25. Такие оценки получены впервые,

поскольку ранее не проводилось масштабных тестов такого рода. Однако отмеченная «зависимость» не является строгой: авторов на букву К примерно в 1,5 раза больше, чем на букву Б, но ошибка у К меньше, чем у Б.

4. Анализ структуры ошибок для метода триграмм

Проведем краткий обзор структуры ошибок для наиболее точного метода идентификации автора – с помощью эталонов авторских триграмм. Детальный анализ полученных результатов планируется провести в дальнейшем, основываясь на предварительной классификации типов ошибок, которые выявились на данном этапе работы.

Несмотря на достаточно большую ошибку на уровне примерно четверти произведений, более 80 тыс. текстов идентифицированы правильно. Это очень важный результат, показывающий, что в основе данного технического метода распознавания автора лежит гипотеза, которая, возможно, не очень сильно отклоняется от главного направления, чем бы это направление ни являлось.

Русскоязычные тексты идентифицированы более точно, чем переводные. Из 66 215 произведений, написанных на русском языке, ошибочно определенных текстов оказалось 12 992, что составило 19,6%. Тексты иностранных авторов были определены с ошибкой 30,1%. По-видимому, это различие следует признать значимым. Возможно, так проявляется «эффект переводчика», приводящий к некоторому искажению авторского текста вследствие того, что переводчики, вообще говоря, разные. Тем не менее, надо учитывать, что наши объяснения этого и последующих эффектов идентификации носят характер домыслов, поскольку и прямо противоположные результаты тоже могут быть объяснены вполне разумно. Например, если бы оказалось, что русскоязычные тексты идентифицированы хуже переводных, можно было бы сказать, что переводились в основном известные профессиональные писатели, которых «переводчиком не испортишь», а русскоязычный самиздат показал недостаток профессионализма. Поэтому приводимые в данном разделе соображения отражают только предположения авторов данной работы.

Интересно было бы провести подробный анализ причин ошибочной идентификации автора каждого текста, однако для такого анализа не хватает объективной информации. Например, довольно большое число авторов (1475) идентифицировано «почти» правильно, ошибочным оказывается только один текст. Возникает вопрос: в силу чего так произошло? Возможно, что автор действительно написал некоторое произведение в несвойственном для себя стиле, причем сделал это сознательно. Тогда наш метод дал бы в определенном смысле правильный результат, не признав автора такого произведения своим. Но возможно, что произведение «типичное», и тогда мы имеем действительно статистическую ошибку метода. Лучше всего на такой вопрос мог бы ответить сам автор, но даже его ответ будет субъективно навязанным. Кроме того, по большому счету, каждое произведение индивидуально, так что не следует

делать и противоположных выводов, подозревая автора в писании своих текстов «под копирку», если он правильно распознается.

Для некоторых авторов объяснение ошибки распознавания может быть дано, как мы считаем, корректно. Например, все художественные произведения Ю. Бондарева определены правильно, за исключением публицистики, которая всегда пишется иначе, чем художественная литература. Аналогично детективные произведения Б. Райнова распознаны точно, а его публицистика – нет. Из других переводных писателей характерны примеры известных авторов детективных романов Рекса Стаута и Дика Фрэнсиса, которые определились полностью, за исключением одного произведения: текста по кулинарии под названием «Поваренная книга Ниро Вульфа» и автобиографической (не детективной) повести «Спорт королев».

С другой стороны, есть и противоположные примеры. У некоторого хорошо распознаваемого автора есть произведение той же серии, что и основные романы, но почему-то идентифицированное как текст другого автора, пишущего примерно в том же жанре. Например, фантастический роман А. Ваклана «Хозяин острова Ниоткуда» идентифицирован как текст, близкий к эталону С. Зверева, а роман А. Мазина «Костер для инквизитора» оказался близок к эталону К. Казанцева. Является ли полученный результат ошибкой метода или он тоже обусловлен некоторой объективной причиной? Под объективной причиной мы подразумеваем не случайную близость выборки к некоторому другому эталону, а намеренное искажение автором своего стиля, в результате чего текст оказывается близким к некоторому другому эталону. Подчеркнем, что здесь мы не предполагаем желание одного из авторов написать нечто в стиле другого автора. Просто метод ближайшего эталона для любого текста предъявит ближайший к нему эталон, даже если правильный эталон отсутствует. Поэтому, если автор написал текст с некоторыми отличиями от своей обычной манеры, то в случае, когда это отклонение повлияло на распределение триграмм, ближайшим эталоном может оказаться эталон другого автора.

Интересным наблюдением является тот факт, что некоторые авторы – как вполне профессиональные литераторы, так и представители «самиздата» – распознаются абсолютно безошибочно. Устойчивость распределения может означать как относительную бедность языка, так и высокий профессионализм. Таких авторов в нашем корпусе оказалось 2191, как русскоязычных, так и в переводах. Опять-таки, с другой стороны, для некоторых авторов методом триграмм не удалось распознать ни одного произведения. Таких авторов 472, среди которых также присутствуют профессиональные писатели и любители. Например, абсолютно не распознались такие известные авторы, как И. Бабель, К. Чапек, М. Зощенко.

Обращает на себя внимание тот факт, что если автор прозы – поэт, то в большинстве случаев его прозаические тексты идентифицируются плохо. Так, абсолютно не распознаны прозаические произведения А. Пушкина, И. Гете, А. Блока, В. Гаршина, И. Бродского, Б. Окуджавы, Ю. Визбора. Перечисленные авторы – не писатели в смысле создания монументальной прозы. Ведь даже

«История Пугачевского бунта» А. Пушкина не вполне произведение художественной литературы. Правда, есть и контрпримеры: К. Бальмонт и М. Цветаева распознаны точно. Возможно, что проблема распознавания отчасти относится не к поэзии, а к журналистике, поскольку, например, Ю. Олеша тоже весь не идентифицирован. В этом плане художественные произведения бывших журналистов (например, И. Ильфа и Е. Петрова) распознаются плохо.

Иногда в качестве объяснения причины ошибки можно выдвинуть гипотезу о двух и более ипостасях автора. Например, все художественные произведения И. Тургенева опознаются точно, кроме некоторых текстов в жанре мистики и близких скорее к белому стиху («Призраки», «Песнь торжествующей любви»). Все тексты Ж. Сименона про Мегрэ опознаются точно, а его другие романы («Президент», «Негритянский квартал», «Братья Рико») не признаются произведениями того же автора. Возможно, это следует считать не ошибкой метода, а правильным «отсутствием распознавания». Но иногда наблюдается и обратная картина. Например, из 70 повестей Н. Лескова правильно опознались лишь 50, а 20 других, написанных вроде бы в том же стиле, ему не соответствуют. То есть не все авторы «удовлетворяют» нашей математической гипотезе, что, впрочем, естественно. В этом плане практически важной задачей является формулировка метода в виде машинного самоконтроля результата применительно хотя бы к некоторым типам ошибок.

Из русских классиков Л. Толстой представляет большую сложность для создания модели коррекции ошибки. В основном его произведения распознаны верно, но ошибки весьма разнородны. Например, явно нравоучительные тексты типа «В чем моя вера» (ближайший эталон для этого произведения – А. Августин) не похожи на «Войну и мир», что, в целом, понятно. Но из трилогии «Детство», «Отрочество», «Юность» средняя часть распознана неверно, она оказалась ближе всего к эталону А. Жида, а это странно. И таких примеров, когда однозначная трактовка ошибки затруднена, для классиков мировой литературы довольно много. Можно было бы попытаться ввести эталоны типа «Толстой-2», «Гоголь-2» и т.п., но жанровая принадлежность текстов не всегда однозначна и, кроме того, не ясно, как определять эту принадлежность в рамках автоматической обработки.

Любопытно отметить, что большинство так называемых «сериальных писателей» (Д. Донцова, А. Маринина, Т. Полякова, Т. Устинова, Н. Леонов и ряд других) показали устойчивость авторского стиля в том смысле, что почти все их произведения распознаются правильно, и потому гипотеза о некоем коллективном творчестве однозначно несостоятельна. Однако обнаружилась и группа из пяти авторов в жанре «боевой фантастики», каждый из которых имеет более ста произведений, причем примерно половина из этих произведений идентифицируется остальными четырьмя членами «могучей кучки». Возможно, эффект объясняется схожестью «маневра», но это единственный пример того, когда можно предположить какие-то прямые взаимосвязи между авторами.

В целом ошибочно идентифицированные тексты группируются по признаку жанра и перевода. То есть переводные любовные романы по большей части (порядка 80%) перепутываются с переводными же романами той же тематики, российская фантастика – с российской фантастикой, философы – с другими философами, а не, допустим, с авторами хорроров. Например, трактаты Аристотеля идентифицировались почти все, а два ошибочно были приписаны Платону и Геродоту. Геродот, в свою очередь, также был распознан абсолютно точно. То, что текст переводной, априори не делает его плохо узнаваемым, потому что, как выяснилось благодаря нашему эксперименту, эффект переводчика влияет не очень сильно, и в большинстве случаев опознается именно автор, а не переводчик. Но надо отметить, что это наблюдение справедливо в основном в тех случаях, когда ошибок распознавания данного автора мало. Если же ошибок много, то они распределены по очень многим жанрам и авторам. Например, Ж.-П. Сартр оказался не распознанным абсолютно весь, и ближайшие к его текстам эталоны не относятся к философам (впрочем, и его тексты тоже не в полном смысле таковы). Аналогичное замечание относится и к произведениям А. Камю и Вольтера. С другой стороны, А. Шопенгауэр и Ф. Ницше распознаны верно, то есть ошибка в данном случае не связана с тематикой текстов.

Также отметим, например, что политическая деятельность автора сама по себе не обязана приводить к хорошему или плохому распознаванию текстов. Произведения Л. Троцкого распознаны верно, а Б. Савинкова – нет.

Таким образом, важным аспектом анализа для понимания способа коррекции ошибки искусственного интеллекта является выяснение реальных причин выявленных отклонений. Косвенные данные, сопровождающие ошибки, на которых может быть обучена система коррекции, не кластеризуются, что требует индивидуального подхода. Причинами могут быть случайный статистический выброс, связанный с невыдержанным стилем автора, специальное авторское целеполагание, скрытые эффекты перевода с других языков.

Представляется, что для правильной интерпретации результатов и разработки модели коррекции ошибки необходимы комментарии экспертов из профессионального сообщества данной направленности, т.е. литераторов, что обычно бывает затруднительно в силу априорного противопоставления искусственного и естественного интеллектов.

Заключение

В данной работе впервые проведен статистический анализ полного корпуса мировой художественной литературы на русском языке методом n-грамм. Исследована структура ошибок распознавания автора по ближайшему эталону в норме L1. Представляется важным сформулировать метод коррекции ошибки для случая, когда автор пишет текст, заведомо не похожий на свой эталон. Тогда правильным ответом является «отсутствие эталона автора в

корпусе». Такая модель была предложена авторами в [11]. Исследование свойств этой модели на полном корпусе будет проведено в отдельной работе.

Разделы 1, 2 и 3 написаны М.Ю. Кислицыной, остальные разделы написаны совместно.

Благодарности

Работа выполнена при поддержке гранта РНФ, проект № 23-71-10055.

Литература

1. Рассел С., Норвиг П. Искусственный интеллект. Современный подход. – М.: Вильямс. 2007. – 1480 с.
2. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. – М.: Вильямс, 2011. – 528 с.
3. Hovy E., Lavid Ju. Towards a science of corpus annotation // International Journal of Translation, 2010. V. 22. No 1. P. 1-25.
4. Шевелев О.Г. Методы автоматической классификации текстов на естественном языке: Учеб. пособие. – Томск: ТМЛ-Пресс, 2007. – 144 с.
5. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
6. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник МГУ. Сер. 9: Филология. 2000. № 2. С. 115 –126.
7. Кукушкина О.В., Поликарпов А.А., Хмелёв Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. 2001. Т. 37. Вып. 2. С. 96 – 109.
8. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы, 2017. Т. 30. № 1. С. 85-99.
9. Кислицын А.А., Кислицына М.Ю. Распознавание выборочных распределений среди системы эталонов: метод ближайшего соседа // Препринты ИПМ им. М.В. Келдыша, 2023. № 11. С. 1-21.
10. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов // Доклады ТУСУРа, № 1, 2009. С. 36-42.
11. Воронина М.Ю., Орлов Ю.Н. Определение автора текста методом сегментации // Компьютерные исследования и моделирование. 2022. Т. 14. № 5. С. 1199-1210.
12. Резанова З.И., Романов А.С., Мещеряков Р.В. О выборе признаков текста, релевантных в автороведческой экспертной деятельности // Вестник Томского государственного университета. Филология. – 2013. – Т. 26. № 6. – С. 38-52.
13. Stamatatos E., Fakotakis N., and Kokkinakis G. Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics, 26(4):471-495, 2000.
14. Argamon S. and Juola P. Overview of the International Authorship Identification Competition at PAN-2011. In V. Petras, P. Forner, P.D. Clough (eds.) CLEF Notebook Papers/Labs/Workshop, 2011.

15. Sudheep E. M., Chinchu J., Puthussery A. and Sasi N. K. Text classification for authorship attribution analysis // Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September 2013.
16. Cappellato L., Ferro N., Halvey M., and Kraaij W. (eds.). CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), 2014.
17. Корпус текстов
https://github.com/akislitsyn/textcorpus (последняя дата обращения 03.03.2024)
18. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1972. – 368 с.
19. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.